投稿至《应用统计年刊》

arXiv:2301.02225v1 [stat.ML] 4 Jan 2023

L_{1-2} GLASSO：L_{1-2} 用于联合估计表达数量性状位点定位和基因网络的 L1−2 正则化多任务图形套索回归

BY WEI MIAO[1,a]，左侧AN YAO[2,b]

[1]C湖南大学数学学院，amiaow@hnu.ed u.cn

[2]湖南大学数学学院，byao@hnu.edu.cn

遗传学中的一个关键问题是揭示细胞内基因表达如何受到调控。调控关联学习的两项主要任务是：（i）识别单核苷酸多态性-基因关系，即表达数量性状位点定位；（ii）确定基因-基因关系，即基因网络估计。为在这两项任务之间共享信息，我们聚焦于联合估计表达数量性状位点定位和基因网络的统一模型，并提出一种 L_{1-2}正则化多任务图形套索回归，命名为 L_{1-2} GLasso。在人工数据集上进行的数值实验表明，L_{1-2} GLasso在捕捉表达数量性状位点定位和基因网络真实稀疏结构方面具有竞争力。 L_{1-2} GLasso进一步应用于ADNI-1真实数据集，实验结果显示，与其他常用方法相比， L_{1-2} GLasso能够获得更稀疏且更准确的解。

1. 引言。自首个人类基因组序列发表以来，测序技术的发展使我们能够获取越来越多的基因组数据。计算技术有助于我们从原始数据中挖掘有意义的信息，并了解细胞中的基因表达如何受到调控。一般而言，这些问题包括识别癌症基因共表达模块（共表达：两个或多个基因同时表达）、通过表达数量性状位点（表达数量性状位点）定位确定单核苷酸多态性-基因关系，以及通过估计基因网络结构确定基因-基因关系等（Rockman和Kruglyak，2006；Gardner和Faith，2005）。给定一个包含单核苷酸多态性（单核苷酸多态性）和mRNA表达的数据集，问题在于理解单核苷酸多态性-基因关系和基因-基因关系。例如，假设有单核苷酸多态性 $x=(x_1,\ldots,x_p)$ 和基因 $y=(y_1,\ldots,y_q)$，表达数量性状位点定位中的单核苷酸多态性-基因关系由回归系数矩阵确定，而基因网络估计中的基因-基因关系则由输出结构捕获。

关于表达数量性状位点定位和基因网络估计，已有许多不同类型的研究。传统的表达数量性状位点定位方法是确定基因与单核苷酸多态性之间是否存在关联。后来，人们开发了多变量模型，用于确定多个单核苷酸多态性与一个基因之间的关系（Michaelson等人，2010）。近年来，又有若干模型被提出，用于确定多个单核苷酸多态性与多个基因之间的关系（Kim 和 Xing，2012年）。

在基因网络估计方面，传统方法是构建一张图，并用一条边连接两个相关基因。具体而言，许多先前研究根据基因表达数据推断基因之间的关系。例如，在高斯图模型（GGM）框架中，图模型使用图来表示随机变量之间的依赖关系（Schäfer and Strimmer，2005；Segalet al.，2005；Li and Gui，2006；Peng et al.，2009）。在 GGM 中，多元向量服从多元正态分布，其协方差矩阵具有特定结构。协方差矩阵的逆称为精度矩阵。GGM 假设，给定基因的表达变异模式可以

关键词和短语：多元回归、结构化稀疏性、 l_1 与 l_2 范数之差。

由其他基因中的一个小子集预测（Meinshausenand Bühlmann，2006）。这一假设使精度矩阵具有稀疏性（即包含多个零），并将问题简化为一个广为人知的邻域选择或协方差选择问题。在精度映射建模框架中，核心思想是使用偏相关作为衡量任意两个基因独立性的指标，从而直接区分直接和间接相互作用。其他方法则利用贝叶斯网络来建立基因之间的结构（Marbach 等人，2010）。

多任务回归模型可用于联合估计回归系数矩阵和输出结构。面临的一项挑战是高维数据灾难，这在遗传数据中十分常见。在以往的研究中，稀疏学习是解决这一问题的有效方法，并因其稀疏解、可解释性强和计算便捷等优势而受到广泛关注（Bertsimas 和 Van Parys，2020年）。此外，为增强表达能力，研究人员提出了各种将稀疏学习与结构化正则化相结合的结构化稀疏模型。在计算和工程的各个领域，基于稀疏性的先验假设以及问题的特定结构特征构建结构化稀疏模型，都是一项重要的研究课题。

已有研究综述了许多基于结构化稀疏正则化的模型（Vinga，2021）。组套索通过促使相关输出组针对相同的输入子集具有非零系数，已得到广泛研究（Yuan 和 Lin，2006 ）。研究人员提出了一种计算效率高的方法，用于对稀疏精度矩阵进行 Lasso 正则化估计（Friedman、Hastie 和 Tibshirani，2008年）。图引导融合套索促使图中相连的输出对具有相似的系数值（Kim 和 Xing，2009）。条件高斯图模型（CGGM）已被用于同时估计输出结构和具有结构化稀疏性的回归系数（Yin 和 Li，2011；Li、Chun 和 Zhao，2012年；Chun 等人，2013年）。然而，这些模型都要求我们预先了解输出 y 之间的关系。另一类模型则侧重于估计 $y|x$ 的条件协方差，而非输出 y 的协方差结构；这类模型已被用于学习回归系数矩阵和输出结构（Rothman、Levina 和 Zhu，2010）。受噪声数据影响，这些模型最终可能无法得到输出之间的真实结构。近期，一种名为逆协方差融合套索（ICLasso）的新方法侧重于输出 y 的协方差结构，也可联合估计回归系数矩阵和输出结构（ Marchetti-Bowick 等人，2019年）。在 ICLasso 中，结构化稀疏正则化惩罚项由 l_1 范数构成。

此外，许多其他正则化惩罚项也得到了研究。为鼓励结构化稀疏模型的稀疏性，研究人员开发了能够同时约束 l_1 和 l_2 范数的投影算子（Hoyer，2004）。其中一种受到广泛研究的正则化惩罚项是 l_1 范数与 l_2 范数之差。该惩罚项被认为具有鲁棒性，有助于选择稀疏解（Yin、Esser 和 Xin，2014年）。它已应用于非负最小二乘（NNLS）和正交匹配追踪。关于成像数据中与 l_1 最小化的比较，可参见（Esser、Lou 和 Xin，2013年）。一些研究人员还在稀疏信号重建问题中应用了 l_1 范数与 l_2 范数之差，以近似原始的基于 l_0范数的稀疏性（Liu 等人，2016年）。在压缩感知、地震反演等其他领域也能看到它的应用（Yin 等人，2015年；Wang 等人，2019年）。

受这些研究的启发，我们提出了一种基于 l_1 与 l_2范数之差的新模型。该模型带来了一些重要的新贡献：(i) 我们在模型中引入了一种新的正则化惩罚项，从而获得更好的近似，并采用更快的算法求解该优化问题。(ii) 在相同的参数设置下，与 MRCE、ICLasso 等现有方法相比，求得的回归系数矩阵更加稀疏。(iii) 在恢复输出结构方面，我们的模型优于其他基线方法。

在第 2 节中，我们介绍几种基线方法。在第 3 节中，我们详细阐述采用不同惩罚项的新模型及其优化算法。在第 4 节和第 5 节中，我们通过模拟数据集和真实数据集评估所提方法的有效性。最后，我们在第 6 节总结全文。

2. 背景。我们假设 $X \in R^{n\times p}$ 是一个 SNP 基因型矩阵， $Y \in R^{n\times q}$是一个基因表达值矩阵。这里， n 表示样本数， q 表示基因数， p 表示单核苷酸多态性数。矩阵的具体形式如下:

$$\underbrace{\begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1q} \\ y_{21} & y_{22} & \cdots & y_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{nq} \end{bmatrix}}_{Y} = \underbrace{\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}}_{X} \times \underbrace{\begin{bmatrix} \beta_{11} & \beta_{12} & \cdots & \beta_{1q} \\ \beta_{21} & \beta_{22} & \cdots & \beta_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{p1} & \beta_{p2} & \cdots & \beta_{pq} \end{bmatrix}}_{B} \tag{1}$$

2.1. 多任务 Lasso。 多任务 Lasso 可用于统计检验，以检测与基因相关的单核苷酸多态性（Tibshirani， 1996）。给定 X 和 Y，多变量线性回归模型表示为

$$y_k = X\beta_k + \epsilon_k, k = 1, \ldots, q, \tag{2}$$

其中， $\beta_k = [\beta_{1k}, \ldots, \beta_{pk}]^T$表示回归系数。它可用于检测与基因显著相关的单核苷酸多态性。我们假设 $\epsilon_k \sim N(0,\ \sigma^2)$，其矩阵形式的数学表达式为:

$$\min_B \frac{1}{n}\|Y - XB\|_F^2 + \lambda\|B\|_1, \tag{3}$$

其中， $B \in R^{p\times q}$ 表示回归系数矩阵， λ 是正则化参数，用于控制稀疏程度。

2.2. 带协方差估计的多元回归。带协方差估计的多元回归（MRCE）是一种能够联合估计回归系数矩阵和输出结构的方法（Rothman、Levina 和 Zhu， 2010）。该方法假设 X与 Y呈线性关系： $Y = XB + E$，其中$E \sim \mathcal{N}\left(0,\ \Omega^{-1}\right)$ 是一个高斯噪声矩阵。我们可以计算得到$Y \mid X \sim \mathcal{N}\left(XB,\ \Omega^{-1}\right)$。MRCE可表示如下:

$$\begin{aligned} \min_{B,\Omega} \frac{1}{n}\operatorname{tr}\left((Y - XB)^T(Y - XB)\Omega\right) \\ - \log\det(\Omega) + \lambda_1\|B\|_1 + \lambda_2\|\Omega\|_1, \end{aligned} \tag{4}$$

其中，Ω 表示 $Y|X$ 的条件逆协方差，而不是 Y的逆协方差。基于先前的假设，Ω 与高斯噪声矩阵 E 相关，无法捕捉回归系数矩阵 B 与 Y中输出结构之间的确切关系。

2.3. 逆协方差融合套索。 Marchetti-Bowick 等人(2019年) 提出了一种称为逆协方差融合套索的新模型。该模型还可以联合估计回归系数和输出结构。与以往研究相比，该方法捕捉的是的边缘逆协方差，而不是的条件逆协方差。

ICLasso（逆协方差融合套索）从两个核心建模假设开始：

$$x \sim \mathcal{N}(0, T)$$
$$y|x \sim \mathcal{N}(x^T B, E),$$

其中 T 表示 x 的协方差， E 表示 $y|x$的条件协方差，Θ 表示 Y的边缘逆协方差，这与Ω 在（4）中不同。

基于这些假设，我们可以推导出 y 的边缘分布。由于边缘分布 p（y）服从高斯分布，因此可以利用全期望定律和全方差定律推导出 y 的均值和协方差，如下所示：

$$\begin{aligned} \mathbb{E}_y(y) &= \mathbb{E}_x\left(\mathbb{E}_{y|x}(y \mid x)\right) = \mathbb{E}_x\left(x^T B\right) = 0 \\ \mathrm{Cov}_y(y) &= \mathbb{E}_x\left(\mathrm{Cov}_{y|x}(y \mid x)\right) + \mathrm{Cov}_x\left(\mathbb{E}_{y|x}(y \mid x)\right) \\ &= \mathbb{E}_x(E) + \mathrm{Cov}_x\left(x^T B\right) = E + B^T T B. \end{aligned} \tag{5}$$

我们可以计算 y 的分布：

$$y \sim \mathcal{N}\left(0, \Theta^{-1}\right), \tag{6}$$

其中， $\Theta^{-1} = E + B^T T B$是 y的边缘协方差。这建立了 y 的逆协方差与回归系数矩阵 B 之间的联系。为简单起见，我们假设$T = \tau^2 I_{p\times p}$和 $E = \varepsilon^2 I_{q\times q}$，因此 $\Theta^{-1} \propto B^T B$。

给定单核苷酸多态性 $x \in R^p$ 和基因 $y \in R^q$ 的独立同分布观测，为了联合估计回归系数矩阵 $B \in R^{p\times q}$ 和逆协方差矩阵 $\Theta \in R^{q\times q}$，逆协方差融合套索优化问题可写为：

$$\begin{aligned} \min_{B,\Theta} \ & \frac{1}{n}\|Y - XB\|_F^2 + \frac{1}{n}\mathrm{tr}\left(Y^T Y \Theta\right) - \log\det(\Theta) \\ & + \lambda_1 \|B\|_1 + \lambda_2 \|\Theta\|_1 \\ & + \gamma \sum_{(k,m)} |\theta_{km}| \cdot \left\|\beta_{.k} + \mathrm{sgn}\left(\theta_{km}\right)\beta_{.m}\right\|_1 . \end{aligned} \tag{7}$$

该目标函数实际上可归结为以下问题的组合：多任务 Lasso 和稀疏逆协方差估计，并包含图引导融合惩罚项。惩罚项 $\sum_{(k,m)} |\theta_{km}| \cdot \|\beta_{.k} + \mathrm{sgn}(\theta_{km})\ \beta_{.m}\|_1$ 旨在促进 B 与 Θ 之间的结构信息共享。

2.4. 使用融合惩罚估计模型参数。先前的研究为我们估计这些参数提供了思路，也促进了 B和 Θ之间的信息共享。为此，该模型被表述为一个凸优化问题，其目标函数为：

$$\mathrm{loss}_{y|x}(B) + \mathrm{loss}_y(\Theta) + \mathrm{penalty}(B, \Theta), \tag{8}$$

其中可以看到：

- 损失$_{y|x}(B)$ 可由 $y|x$ 的负对数似然推导得出；
- 损失$_y(\Theta)$ 可由 y 的负边际对数似然推导得出；
- 惩罚项(B,Θ) 是一种惩罚项，用于促进回归系数矩阵 B 与输出结构 Θ 之间的信息共享。

l_1 范数惩罚项 $\|B\|_1$ 和 $\|\Theta\|_1$ 使 B 和 Θ的估计具有稀疏性，从而使模型即使在高维数据上也可行。为了获得更稀疏的解，我们可以自然地推广 l_1 范数惩罚项。基于这一框架，我们将在下一节提出我们的模型。

3. $\boldsymbol{l_{1-2}}$ 图形套索。 由于 B 的稀疏性体现为其非零项的数量，因此这等价于所谓的 l_0 范数惩罚项。ICLasso 将 l_0 范数惩罚项替换为 l_1 范数惩罚项。Esser、Lou 和 Xin（Esser、Lou 和 Xin，2013年）提出了一种基于差值 l_1 和 l_2范数的重构框架。该框架可改写如下：

$$\begin{aligned} &\min_x \lambda\|x\|_1-\tau\|x\|_2 \\ &\text{s.t.} \quad y=Ax. \end{aligned} \tag{9}$$

已有证明表明，只要选择适当的 λ 和 τ，所得解就可以接近采用 l_0 范数惩罚项的问题的解。为了获得更稀疏的解，我们提出了l_{1-2} 图形套索。我们的模型可描述为：

$$\text{loss}_{y|x}(B)+\text{loss}_y(\Theta)+\text{penalty}(B,\Theta)+\text{penalty}(B)+\text{penalty}(\Theta). \tag{10}$$

具体而言，表达式如下：

$$\begin{aligned} \min_{B,\Theta}\ & \frac{1}{n}\|Y-XB\|_F^2+\frac{1}{n}\operatorname{tr}\left(Y^TY\Theta\right)-\log\det(\Theta) \\ & +\lambda_1||B||_1-\tau||B||_{2,1}+\lambda_2\|\Theta\|_1 \\ & +\gamma\sum_{(k,m)}|\theta_{km}|\cdot\left\|\beta_{\cdot k}+\operatorname{sgn}\left(\theta_{km}\right)\beta_{\cdot m}\right\|_1. \end{aligned} \tag{11}$$

我们定义：

$$g_{12}(B)=\frac{1}{n}||Y-XB||_F^2+\lambda_1||B||_1-\tau||B||_{2,1} \tag{12}$$

$$h(\Theta)=\frac{1}{n}tr(Y^TY\Theta)-\log\det(\Theta)+\lambda_2||\Theta||_1 \tag{13}$$

$$\text{GFL}(B,-\Theta)=\sum_{k=1}^{q}\sum_{m=1}^{q}|\theta_{km}|\cdot\left\|\beta_{\cdot k}+\operatorname{sgn}\left(\theta_{km}\right)\beta_{\cdot m}\right\|_1 \tag{14}$$

项$\frac{1}{n}\|Y-XB\|_F^2$，$\frac{1}{n}\operatorname{tr}\left(Y^TY\Theta\right)$–$\log\det(\Theta)$和$\gamma\sum_{(k,\ m)}|\theta_{km}|\cdot\|\beta_k+\operatorname{sgn}\left(\theta_{km}\right)\beta_{\cdot m}\|_1$分别由公式（8）推导得出。我们详细说明每一项的作用：

- 损失$_{y|x}(B)$：$\frac{1}{n}\|Y-XB\|_F^2$。根据 $y\mid x\sim\mathcal{N}\left(x^TB,\ \varepsilon^2I\right)$，我们可以通过最大似然估计推导出其表达式。该项的作用是促使回归系数矩阵 B。
- 损失$_y(\Theta)$:$\frac{1}{n}\operatorname{tr}\left(Y^TY\Theta\right)-\log\det(\Theta)$。根据 $y\sim\mathcal{N}\left(0,\ \Theta^{-1}\right)$，我们可以通过最大似然估计推导其表达式。该项旨在促使逆协方差 Θ 反映输出之间的相关性。

• 惩罚项(B, Θ): $\gamma \sum_{(k,\ m)} |\theta_{km}| \cdot \|\beta_{\cdot k} + \operatorname{sgn}(\theta_{km})\, \beta_{\cdot m}\|_1$ 是一种图引导融合惩罚。它促使关系密切的输出对应的回归系数彼此相似。当 y_k 与 y_m 之间呈部分正相关，且对于任意 j， $\beta_{jk} \neq \beta_{jm}$ 时，它会施加与 θ_{km}成正比的惩罚；当对于任意j， y_k 和 y_m 之间存在负偏相关时，它会施加与 $-\theta_{km}$成正比的惩罚。

• 惩罚项(B): $\lambda_1||B||_1 - \tau||B||_{2,\ 1} = \lambda_1 \sum_{j,\ k} |\beta_{jk}| + \tau_2 \sum_{j=1}^{p} \sqrt{\sum_{k=1}^{q} \beta_{jk}^2}$ 是作用于回归系数矩阵的 l_{1-2}范数惩罚项，可在 B中诱导稀疏性。与 l_1 范数相比， l_1 范数和 l_2 范数之差更接近 l_0 范数。

• 惩罚项(Θ): $\lambda_2 \|\Theta\|_1 = \lambda_2 \sum_{k,\ m} |\theta_{km}|$ 是可在 Θ中诱导稀疏性的 l_1 范数惩罚项。

3.1. 与 ICLasso 的关系。l_1-2-图形套索还隐含地假设了两个基本建模假设： $x \sim \mathcal{N}(0,\ T)$ 和$y|x \sim \mathcal{N}(x^T B,\ E)$。$l_1$-2-图形套索与 ICLasso 的区别在于，我们使用 l_1 范数与 l_2 范数之差来获得更接近现实世界的稀疏解。实验结果将在下一节中展示。我们发现，与其他模型相比，l_1-2-图形套索在保持回归误差的同时，可以获得更稀疏的解。

3.2. 优化。 关于该问题的先前研究（11）提出了一些现成的算法来求解上述优化问题。基于这些研究，我们采用交替最小化策略求解 l_1-2-图形套索。注意 GFL(B, $-\Theta$) 项，显然该项是一个双凸函数。因此，定义

$$
\begin{aligned}
g_{12}(B) &= \frac{1}{n}\|Y - XB\|_F^2 + \lambda_1 \|B\|_1 - \tau||B||_{2,1} \\
h(\Theta) &= \frac{1}{n} \operatorname{tr}\left(Y^T Y \Theta\right) - \log\det(\Theta) + \lambda_2 \|\Theta\|_1.
\end{aligned} \tag{15}
$$

原始目标函数可以重写为

$$
\min_{B,\Theta}\ g_{12}(B) + h(\Theta) + \mathrm{GFL}(B, -\Theta). \tag{16}
$$

与 ICLasso 相比，l_1-2-图形套索通过 l_1 范数和 l_2 范数之差，可以得到更稀疏的回归系数矩阵。求解该问题的难点也在于这一惩罚项。可以看出，项 GFL($B, -\Theta$) 是双凸的，因此我们可以采用 Marchetti-Bowick 等人 (2019年) 提出的交替最小化策略来求解问题（15）。

首先，固定 Θ，则问题变为：

$$
f_\Theta(B) = g_{12}(B) + \mathrm{GFL}(B, -\Theta). \tag{17}
$$

尽管我们的模型引入了一个新的正则化项 $\lambda_1||B||_1 - \tau||B||_{2,\ 1}$，但它可以分解为可微函数和不可微函数的形式。对于矩阵 $B = [\boldsymbol{\beta_1}, \boldsymbol{\beta_2}, \ldots, \boldsymbol{\beta_p}]^T$， $\boldsymbol{\beta_i}^T$是 B 中的第 ith 行。根据定义：

$$
||B||_{2,1} = \sum_{i=1}^{p} (\boldsymbol{\beta_i}^T \boldsymbol{\beta_i})^{\frac{1}{2}}, \tag{18}
$$

我们定义 Σ 以及$||B||_{2\text{的}1}$ 导数：

$$
\Sigma := \begin{bmatrix}
\frac{1}{||\boldsymbol{\beta_1}||_2} & 0 & \cdots & 0 \\
0 & \frac{1}{||\boldsymbol{\beta_2}||_2} & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & \frac{1}{||\boldsymbol{\beta_p}||_2}
\end{bmatrix}, \tag{19}
$$

$$\frac{\partial||B||_{2,1}}{\partial B} = \left(\frac{\partial \sum_{i=1}^{p} (\boldsymbol{\beta_i}^T \boldsymbol{\beta_i})^{\frac{1}{2}}}{\partial \boldsymbol{\beta_i}}\right)_{p\times 1} = \begin{bmatrix} \frac{1}{||\boldsymbol{\beta_1}||_2} & 0 & \cdots & 0 \\ 0 & \frac{1}{||\boldsymbol{\beta_2}||_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{||\boldsymbol{\beta_p}||_2} \end{bmatrix} B = \Sigma B. \tag{20}$$

当我们使用较小的 τ 时，项 $\|Y - XB\|_F^2 - \tau||B||_{2,\ 1}$ 被视为凸函数。这个问题可以通过近端平均近端梯度下降（PA-PG）算法Yu（2013年）求解。与近端梯度下降相比，PA-PG 的收敛速度始终更快。已有证明表明，对于任意精度 $O(1/\epsilon)$，在合适的步长下，次梯度法至多经过 $\epsilon > 0$ $\|Y - XB\|_F^2 - \tau||B||_{2\ 1}$ $X^T\ XB -$ 步即可 $Y\ -\tau\Sigma$ $B - \nu\ X^T\ XB - Y\ -\tau\Sigma$ 收敛。首先，的推导过程为（），然后采用形式为（（））的梯度步，并使用一些较小的步长 ν。其他优化过程可参照Marchetti-Bowick 等人（2019年）。最后，这个子问题可写为

$$\widetilde{\beta}_{jk}, \widetilde{\beta}_{jm} = \arg\min_{\beta_{\mathrm{jk}}, \beta_{\mathrm{jm}}} \frac{1}{2\nu}(\beta_{\mathrm{jk}} - \mathrm{z_{jk}})^2 + (\beta_{\mathrm{jm}} - \mathrm{z_{jm}})^2 + |\beta_{\mathrm{jk}} + \mathrm{sgn}(\theta_{\mathrm{km}})\beta_{\mathrm{jm}}|. \tag{21}$$

我们可以求得 β 的闭式解。这个子问题的解相对简单，算法的收敛速度也可见于 Yu (2013年)。

然后固定 B，问题变为：

$$\min_{\Theta}\ h(\Theta) + \mathrm{GFL}(B, -\Theta). \tag{22}$$

这个问题可以通过调整块坐标下降（BCD）算法来求解（Friedman、Hastie 和 Tibshirani，2008年）。最后，这个子问题可以写为：

$$\min_{\alpha}\ \frac{1}{2}\alpha^T \widetilde{H}_j \alpha + u^T \alpha_+ - l^T \alpha_-. \tag{23}$$

我们使用坐标下降法求解每个坐标，并应用软阈值算子的一个变体。

4. 模拟研究。本节中，我们在已知 B 和 Θ 取值的合成数据上比较不同模型，从而直接衡量真实参数值的恢复效果。比较的模型包括图引导融合套索（GFLasso）、带协方差估计的稀疏多元回归（MRCE）、逆协方差融合套索（ICLasso）以及 l_1-2-图形套索。对于每个模型，我们通过最小化留出验证集上的误差来选择超参数值。

我们从两个维度评估每个模型：（i）稀疏结构的恢复。（ii） B 和 Θ的回归误差。

（i）稀疏结构的恢复：为了评估每个模型对 B 和 Θ稀疏结构的估计效果，我们计算恢复 B和 Θ元素的F1分数。为此，我们选择一个阈值，将各元素的值判定为“零”或“非零”，然后计算不同合成数据下的精确率（P）和召回率（R）。

（ii） B 和 Θ的回归误差：我们考察每个模型在ˆ使用 $\boldsymbol{B}$ 根据 X预测 Y时的预测误差。

在这项分析中，我们在输出上使用块结构网络。输出被划分为互不重叠的组，每组在网络中构成一个完全连接的子图。

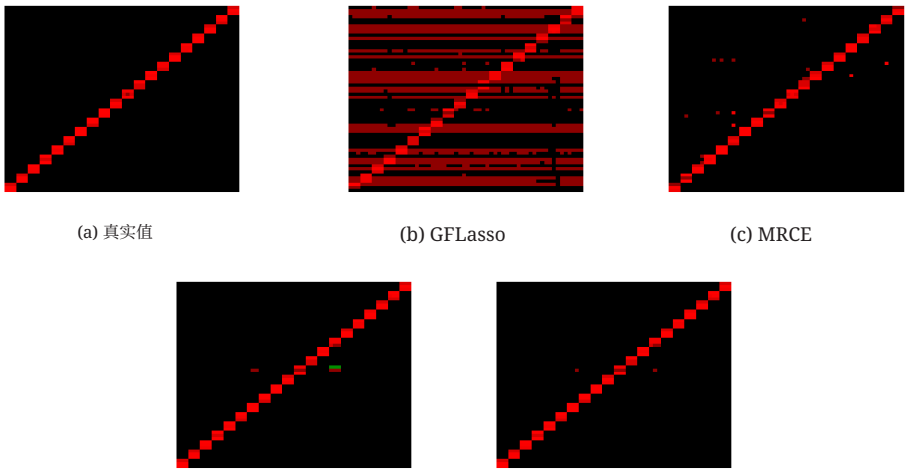
这里我们介绍生成合成数据的流程。总体而言，我们首先固定模型底层组成部分的稀疏结构，然后生成系数值，最后根据模型对 X 和 Y 进行采样。

给定样本数 n、基因数 q 和单核苷酸多态性数 p，首先确定基因网络中的模块大小，并固定与每个基因相关的单核苷酸多态性数 s。接下来，将每个基因随机分配到一个模块，并选取与每个模块相关的 s 个单核苷酸多态性的集合。对于每个模块，我们随机指定其中一个基因为主基因；对于与主基因 y_k 相关的每个单核苷酸多态性 x_j，根据 $\beta_{jk} \sim Uniform(0, 1)$生成其关联强度。对于该模块中的其他基因 y_m，则根据 $\beta_{jm} \sim Uniform(\beta_{jk}, \rho^2)$生成关联强度，其中 $\rho = 0.1$（我们可以针对不同的合成数据更改此参数）。

接下来，我们考虑协方差矩阵 E 和 T 的四种设置，以生成模拟数据集。它们分别是：情形1：$T = I_{pp}$ 和 $E = I_{qq}$。情形2：$T = 0.6^{|j-k|}$ 和 $E = I_{qq}$。情形3：$T = I_{pp}$ 和 $E = 0.6^{|j-k|}$。情况 4：$T = 0.6^{|j-k|}$ 和 $E = 0.6^{|j-k|}$。最后，根据 (5)，我们生成 $\Theta = (E + B^T T B)^{-1}$。对于每种情形，我们对 15 个合成数据集上的结果取平均值。

在一个包含 $n = 120, p = 60, q = 60$ 的合成数据集上的结果比较见图 1、2。具体描述如下：我们确定了 20 个组，每组包含 3 个基因。在每个组中，y_1 与 x_1、x_2、x_3 相关；y_4 与 x_4、x_5、x_6 相关，依此类推。令 $E = I_{qq}$ 和 $T \neq I_{pp}$。

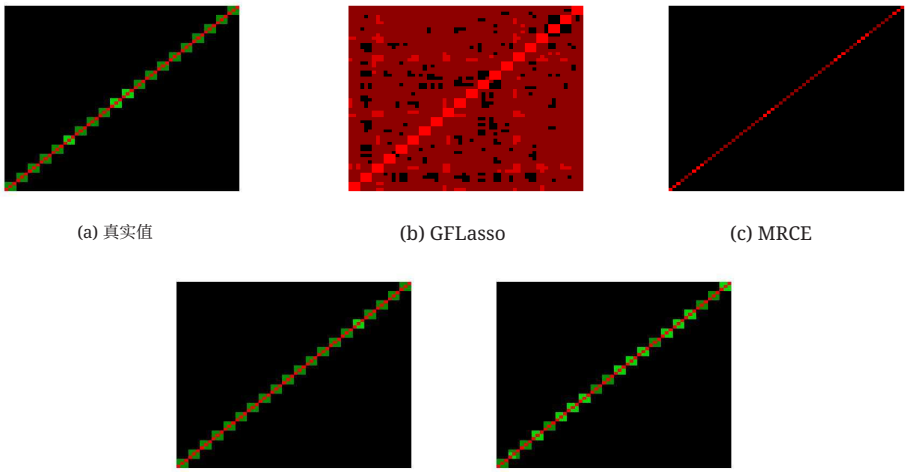
真实的 B 和 Θ 均显示在图1、2的左上角；图1、2 的右侧展示了 GFLasso 和 MRCE 的结果，底部一行展示了 ICLasso 和 $l_{1\text{-}2}$-图形套索的结果。所有模型都能找到 B的结构，但 GFLasso 的估计存在显著误差。与其他模型相比，$l_{1\text{-}2}$-图形套索计算出的回归系数矩阵更接近真实的 B。



(a) 真实值　(b) GFLasso　(c) MRCE

(d) ICLasso　(e) $l_{1\text{-}2}$-图形套索

FIG 1. B 与单个合成数据集上的不同模型的比较（$p = 60$ 和 $q = 60$）。真实的 B 位于图的左上角。

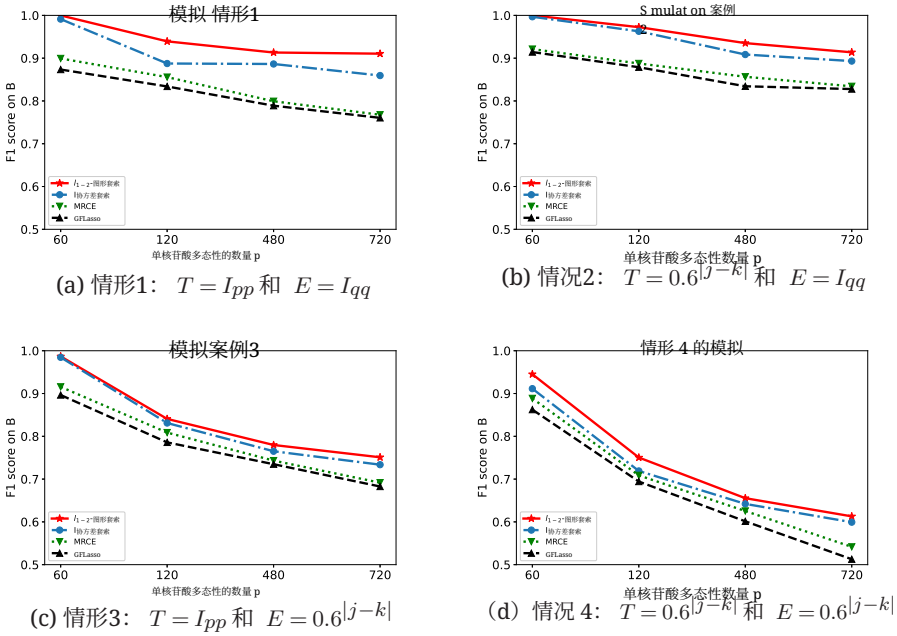


(a) 真实值 (b) GFLasso (c) MRCE

(d) ICLasso (e) $l_{1\text{-}2}$-图形套索

FIG 2. 在单个合成数据集（$p = 60$ 和 $q = 60$）上比较 Θ 在不同模型下的结果。真实的 Θ 位于图的左上角。

从图2可以看出，只有$l_{1\text{-}2}$-图形套索和 ICLasso 能够更准确地恢复 Θ 。合成实验的主要结果如图3、4、5所示。我们根据三项指标评估所提出的方法：(1) B 上的 F1 分数；(2) Θ 上的 F1 分数；(3) Y上的回归误差。可以看出，在稀疏性方面，$l_{1\text{-}2}$-图形套索更接近真实数据。比较不同输入维度的大小后，$l_{1\text{-}2}$-图形套索也优于其他模型。



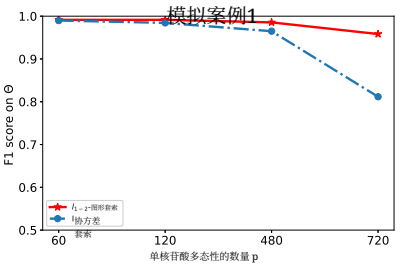
(a) 情形1： $T = I_{pp}$ 和 $E = I_{qq}$

(b) 情况2： $T = 0.6^{|j-k|}$ 和 $E = I_{qq}$

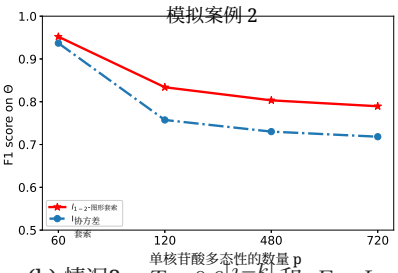
(c) 情形3： $T = I_{pp}$ 和 $E = 0.6^{|j-k|}$

(d) 情况 4： $T = 0.6^{|j-k|}$ 和 $E = 0.6^{|j-k|}$

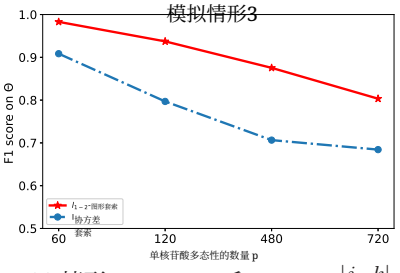
图 3 。在 15 个采用不同类型协方差结构生成的合成数据集上对 B 进行比较（ T, E）一个 d，包含不同数量的单核苷酸多态性 p。



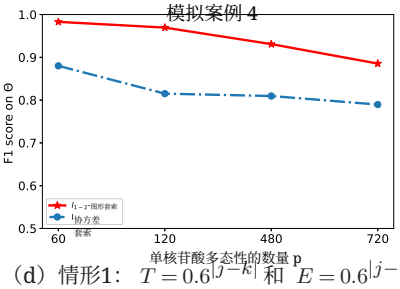
（a）情形1： $T = I_{pp}$ 和 $E = I_{qq}$



(b) 情况2： $T = 0.6^{|j-k|}$ 和 $E = I_{qq}$

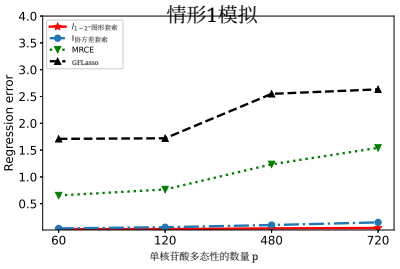


(c) 情形3： $T = I_{pp}$ 和 $E = 0.6^{|j-k|}$

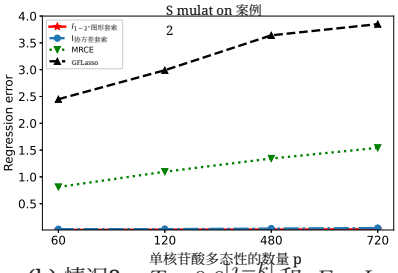


（d）情形1： $T = 0.6^{|j-k|}$ 和 $E = 0.6^{|j-k|}$

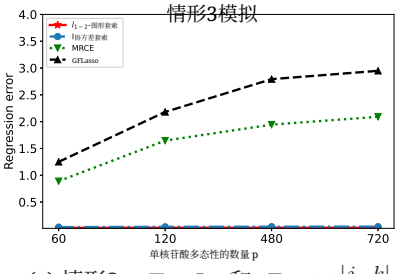
F图 4。在 15 个由不同类型的协方差结构（T、E ）生成的合成数据集上，对 Θ 进行比较，且单核苷酸多态性的数量不同 p。



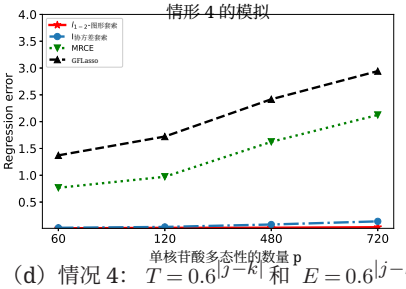
(a) 情形1： $T = I_{pp}$ 和 $E = I_{qq}$



(b) 情况2： $T = 0.6^{|j-k|}$ 和 $E = I_{qq}$



(c) 情形3： $T = I_{pp}$ 和 $E = 0.6^{|j-k|}$



（d）情况 4： $T = 0.6^{|j-k|}$ 和 $E = 0.6^{|j-k|}$

F图 5. 在 15 个具有不同类型协方差结构（T、 E）和不同单核苷酸多态性数量 p 的合成数据集上，对回归误差进行比较。

在图 3、4中，我们展示了恢复 B 和 Θ的 F1分数。在某种程度上，F1分数反映了各模型学习回归系数矩阵和

输出结构。在图 5中，我们展示了 Y的回归误差。从回归误差来看，$l_{1\text{-}2}$-图形套索可能略优于 ICLasso。正如这些图所示，我们的结果清楚地表明，在所考虑的四种情况下，$l_{1\text{-}2}$-图形套索优于其他基线方法。

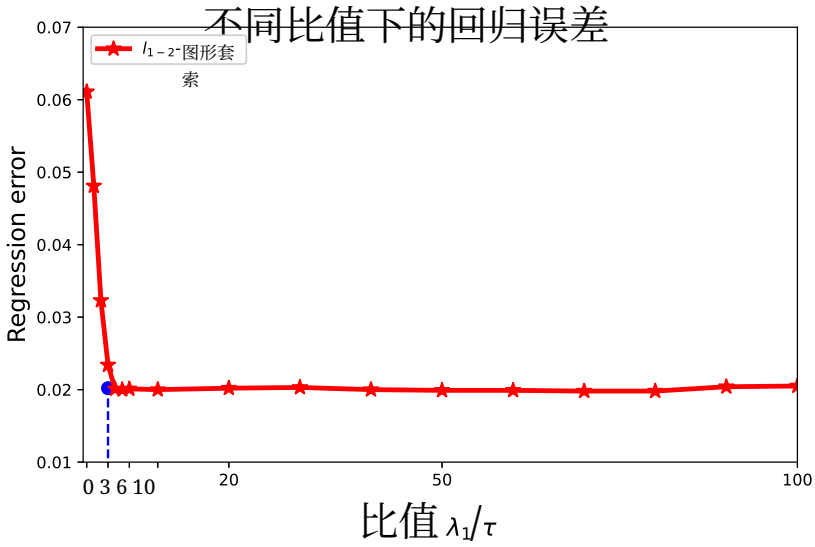


图6。不同比率下的回归误差 $\frac{\lambda_1}{\tau}$

最后，由于 $l_{1\text{-}2}$-GLasso 引入了一个新的超参数 τ，我们还研究了不同超参数对模型效果的影响。我们使用图 1所示的数据集，并在图 $\frac{\lambda_1}{\tau}$中展示了不同比率下的回归误差 6。可以看到，当 $\frac{\lambda_1}{\tau} < 3$时，回归误差最小。较大的比率会带来更令人满意的解。当 $\frac{\lambda_1}{\tau} \geq 3$时，回归误差变得越来越稳定，因此我们最终选择 $\frac{\lambda_1}{\tau} = 10$ 来开展后续实验。

5. 来自表达数量性状位点定位的真实数据集。Greenlaw 等人（2017）分析了一个从 ADNI-1数据库获取的数据集。我们在该数据集上比较$l_{1\text{-}2}$-图形套索与其他模型。我们分析中使用的基因列于表1。

TABLE 1 ADNI-1数据库中的基因标识符

基因标识符	测量指标	感兴趣区域
左侧_杏仁核体积	体积	杏仁核
Left_小脑皮层	体积	大脑皮层
Left_小脑白质	体积	大脑白质
左侧_海马体积	体积	海马体
Left_下外侧脑室	体积	下外侧脑室
Left_侧脑室	体积	侧脑室
Left_内嗅皮层	厚度	内嗅皮层
左侧梭状回_	厚度	梭状回
左侧下顶叶_	厚度	下顶叶回
左侧_下颞回	厚度	下颞回
左侧_中颞回	厚度	中颞回
左侧_海马旁回	厚度	海马旁回
左侧_后扣带回	厚度	后扣带
左侧中央后回_	厚度	中央后回
左侧中央前回_	厚度	中央前回

该数据集包含 $n = 632$ 名受试者；在所有可能的单核苷酸多态性中，我们仅纳入AlzGene数据库所列排名前 15 位候选基因中的单核苷酸多态性。此处所示数据集查询自ADNI-1数据库中截至2014年12月最新的基因组版本。经过质量控制和基因型填补后，本研究使用的遗传数据集包含来自 $q = 15$ 个基因的 $p = 486$ 个单核苷酸多态性。

我们应用$l_{1\text{-}2}$-图形套索及其他模型，来发现影响基因表达水平的单核苷酸多态性。在遗传学界，这类研究通常称为表达数量性状位点（eQTL）定位。人们普遍认为，当基因组中的遗传变异（如单核苷酸多态性）扰动某个基因的表达时，这种影响会沿基因网络传播，进而影响通路下游基因的表达。首先，我们使用480个样本估计回归系数矩阵，然后在剩余的152个样本上计算回归误差。如图7所示，$l_{1\text{-}2}$-图形套索的回归误差最小。需要注意的是，我们每次随机选取480个样本，并重复抽样10次。我们展示了这10个不同数据集上的回归误差中位数（红色）及其离散分布。

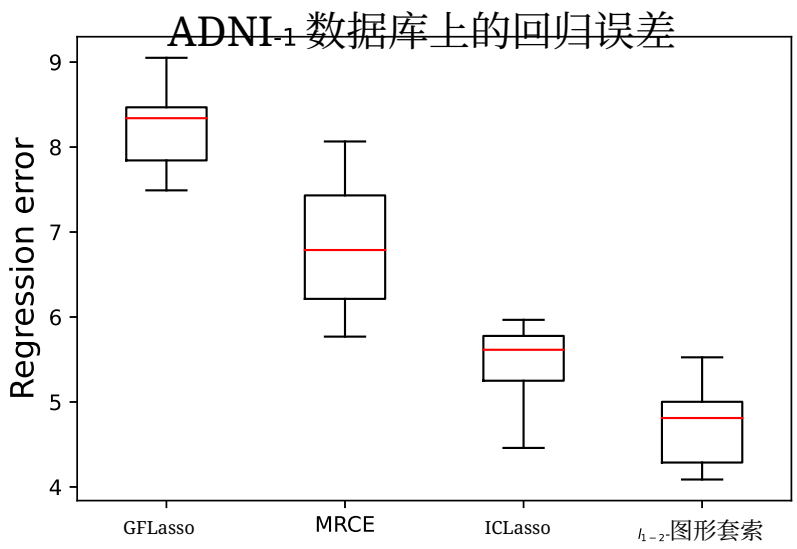
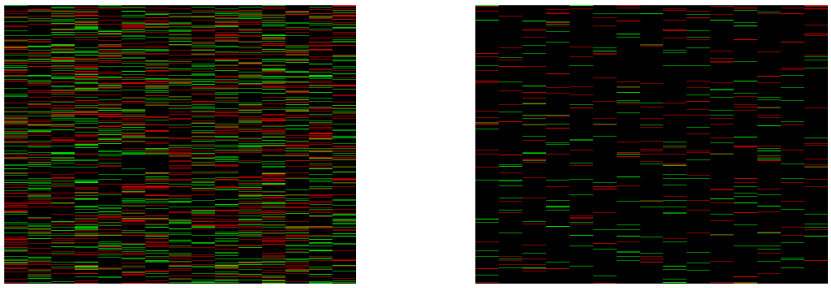


图7。每个箱线图展示了这 10 个不同数据集上回归误差的离散分布。红线表示中位数。

在 Wang 等人（2012年）提出的原始方法中，基于生物学实验，检测到了与15个基因高度相关的24个单核苷酸多态性。根据生物学实验结果，我们也应用所提出的模型来寻找相关的单核苷酸多态性。$l_{1\text{-}2}$-图形套索在已确定的24个单核苷酸多态性中找到了22个，这22个位点在表5中以突出形式标出。此外，我们还在表2、3、4中将数据集上的GFLasso、MRCE和ICLasso与我们的模型进行了比较。各模型选出的单核苷酸多态性以粗体标示。从表4、5可以看出，$l_{1\text{-}2}$-图形套索和ICLasso均能估计基因网络中两个较大子网络的结构。生物学实验已证实这一点。实验表明，两个单核苷酸多态性：rs405509和rs10787010尤为突出，它们可能与数量最多的感兴趣区域相关。在表6中，我们展示了不同模型识别出的与基因相关的单核苷酸多态性。$l_{1\text{-}2}$-图形套索优于其他模型。为了展示$l_{1\text{-}2}$-图形套索的优势，我们可以看到，从表2、3、4可知，GFLasso找出了已确定的24个单核苷酸多态性中的14个，MRCE找出了16个，ICLasso找出了17个。实验结果与我们对$l_{1\text{-}2}$-图形套索的假设一致。与其他模型相比，$l_{1\text{-}2}$-图形套索能够更准确地确定与基因相关的单核苷酸多态性。

回到稀疏性问题，我们将 B 由$l_{1\text{-}2}$- 图形套索计算出的回归系数矩阵与 ICLasso 进行比较。如图8所示，所提出的模型能够得到更稀疏的解。例如，我们重点考察单核苷酸多态性 rs4311。在真实数据集中，研究发现单核苷酸多态性 rs4311 与基因顶下小叶（左侧）存在关联（Wang 等人，2012年）。ICLasso 不仅发现单核苷酸多态性 rs4311 与基因顶下小叶（左侧）之间存在关联，还检测到该单核苷酸多态性与基因左侧杏仁核体积、左侧小脑皮层、左侧侧脑室、左侧内嗅皮层和左侧梭状回相关。而在我们的模型中，$l_{1\text{-}2}$-图形套索计算出的单核苷酸多态性 rs4311 与基因左侧杏仁核体积、左侧小脑皮层、左侧侧脑室、左侧内嗅皮层和左侧梭状回之间的系数分别为 0.019947、0、0、0、0。对于与某个单核苷酸多态性无关的基因，$l_{1\text{-}2}$-图形套索计算出的回归系数大多为0。结果表明，与 ICLasso 相比，我们的模型能够得到更稀疏的解。



(a) B 由 ICLasso 计算　　(b) B 由 $l_{1\text{-}2}$-图形套索计算

图8 展示了由 ICLasso 和 $l_{1\text{-}2}$-图形套索计算得到的回归系数矩阵。$l_{1\text{-}2}$-图形套索能够得到更稀疏、更准确的回归系数矩阵 B

.

表2 生物学实验中识别出的 24 个单核苷酸多态性位点中，包含使用 GFLasso 选出的 14 个单核苷酸多态性位点

SNP	表型数量	表型编号（半球）
rs4311	1	顶下小叶（左侧）
rs405509	8	左侧杏仁核体积、左侧小脑白质、左侧梭状回、海马体积（左侧）、顶下小叶（左侧）、左侧下颞回、左侧中颞回、左侧中央后回
rs666004	1	左侧下颞回
rs1433099	1	左侧小脑皮层
rs1473180	4	左侧小脑皮层、左侧内嗅皮层、左侧梭状回、左侧后扣带回
rs1475345	1	左侧海马旁回
rs1568400	1	左侧中央前回
rs2149196	1	左侧中央后回
rs2418811	1	左侧小脑白质
rs4935774	1	左侧小脑白质
rs6107516	1	左侧中颞回
rs6584307	1	左侧海马旁回
rs11191692	1	左侧内嗅皮层
rs12209631	2	左侧小脑皮层，左侧海马体积
rs16924159	1	左侧后扣带回
rs1023024	1	左侧中央前回
rs1269918	3	左侧小脑皮层、左侧小脑白质、左侧下外侧脑室
rs2327389	1	左侧杏仁核体积
rs2418811	1	左侧小脑白质
rs2756271	4	左侧内嗅皮层、左侧海马体积、左侧下颞回、左侧海马旁回
rs7219773	1	左侧中央前回
rs9314349	1	左侧海马旁回
rs10787010	7	左侧杏仁核体积、左侧内嗅皮层、左侧梭状回、左侧海马体积，左侧下外侧脑室、左侧下颞回、左侧中央前回
rs10787011	1	左侧内嗅皮层
rs11601726	2	左侧小脑白质，左侧侧脑室

表3 生物学实验中鉴定出的24个单核苷酸多态性中，有16个是使用MRCE选出的

SNP	表型数量	表型编号（半球）
rs4311	1	顶下小叶（左侧）
rs405509	3	左侧杏仁核体积、左侧小脑白质、左侧海马体积
rs666004	1	左侧下颞回
rs1433099	1	左侧小脑皮层
rs1473180	4	左侧小脑皮层，左侧内嗅皮层、左侧梭状回、左侧后扣带回
rs1475345	1	左侧海马旁回
rs1568400	1	左侧中央前回
rs2149196	1	左侧中央后回
rs2327389	1	左侧杏仁核体积
rs4935774	1	左侧小脑白质
rs6107516	1	左侧中颞回
rs6584307	1	左侧海马旁回
rs9314349	1	左侧海马旁回
rs10787010	2	左侧下外侧脑室，左侧中央前回
rs11191692	1	左侧内嗅皮层
rs12209631	2	左侧小脑皮层，左侧海马体积
rs1023024	1	左侧中央前回
rs1269918	3	左侧小脑皮层，左侧小脑白质，左侧下外侧脑室
rs2418811	1	左侧小脑白质
rs2756271	4	左侧内嗅皮层、左侧海马体积、左侧下颞回、左侧海马旁回
rs7219773	1	左侧中央前回
rs10787011	1	左侧内嗅皮层
rs11601726	2	左侧小脑白质、左侧侧脑室
rs16924159	1	左侧后扣带回

表4 生物学实验中鉴定出的 24 个单核苷酸多态性包含使用 ICLasso 选出的 17 个单核苷酸多态性

SNP	表型数量	表型编号（半球）
rs4311	1	顶下小叶（左侧）
rs405509	8	左侧杏仁核体积，左侧小脑白质，左侧梭状回，左侧海马体积，顶下小叶（左侧），左侧下颞回，左侧中颞回，左侧中央后回
rs1433099	1	左侧小脑皮层
rs1473180	4	左侧小脑皮层，左侧内嗅皮层，左侧梭状回，左侧后扣带回
rs1475345	1	左侧海马旁回
rs1568400	1	左侧中央前回
rs2149196	1	左侧中央后回
rs2418811	1	左侧小脑白质
rs4935774	1	左侧小脑白质
rs6107516	1	左侧中颞回
rs6584307	1	左侧海马旁回
rs9314349	1	左侧海马旁回
rs10787010	7	左侧杏仁核体积、左侧内嗅皮层、左侧梭状回、左侧海马体积，左侧下外侧脑室、左侧下颞回、左侧中央前回
rs10787011	1	左侧内嗅皮层
rs11191692	1	左侧内嗅皮层
rs12209631	2	左侧小脑皮层，左侧海马体积
rs16924159	1	左侧后扣带回
rs666004	1	左侧下颞回
rs1023024	1	左侧中央前回
rs1269918	3	左侧小脑皮层、左侧小脑白质、左侧下外侧脑室
rs2327389	1	左侧杏仁核体积
rs2756271	4	左侧内嗅皮层、左侧海马体积、左侧下颞回、左侧海马旁回
rs7219773	1	左侧中央前回
rs11601726	2	左侧小脑白质、左侧侧脑室

表5 生物学实验中鉴定的 24 个单核苷酸多态性包含使用 -2-图形套索选出的 22 个单核苷酸多态性

SNP	表型数量	表型编号（半球）
rs4311	1	顶下小叶（左侧）
rs405509	8	左侧杏仁核体积、左侧小脑白质、左侧梭状回、海马体积（左侧）、顶下小叶（左侧）、左侧下颞回、左侧中颞回、左侧中央后回
rs666004	1	左侧下颞回
rs1023024	1	左侧中央前回
rs1269918	3	左侧小脑皮层、左侧小脑白质、左侧下外侧脑室
rs1433099	1	左侧小脑皮层
rs1473180	4	左侧小脑皮层，左侧内嗅皮层，左侧梭状回，左侧后扣带回
rs1475345	1	左侧海马旁回
rs1568400	1	左侧中央前回
rs2149196	1	左侧中央后回
rs2418811	1	左侧小脑白质
rs2756271	4	左侧内嗅皮层、左侧海马体积、左侧下颞回、左侧海马旁回
rs4935774	1	左侧小脑白质
rs6107516	1	左侧中颞回
rs6584307	1	左侧海马旁回
rs9314349	1	左侧海马旁回
rs10787010	7	左侧杏仁核体积、左侧内嗅皮层、左侧梭状回、左侧海马体积，左侧下外侧脑室、左侧下颞回、左侧中央前回
rs10787011	1	左侧内嗅皮层
rs11191692	1	左侧内嗅皮层
rs11601726	2	左侧小脑白质，左侧侧脑室
rs12209631	2	左侧小脑皮层，左侧海马体积
rs16924159	1	左侧后扣带回
rs2327389	1	左侧杏仁核体积
rs7219773	1	左侧中央前回

TABLE 6 个模型识别出的单核苷酸多态性

SNP	GFLasso	MRCE	ICLasso	$l_{1\text{-}2}$-图形套索
rs4311	✘	✘	✘	✘
rs405509	✘	✘	✘	✘
rs666004		✘		✘
rs1023024				✘
rs1269918				✘
rs1433099	✘	✘	✘	✘
rs1473180	✘	✘	✘	✘
rs1475345	✘	✘	✘	✘
rs1568400	✘	✘	✘	✘
rs2149196	✘	✘	✘	✘
rs2327389		✘		
rs2418811	✘		✘	✘
rs2756271				✘
rs4935774		✘	✘	✘
rs6107516	✘	✘	✘	✘
rs6584307	✘	✘	✘	✘
rs7219773				
rs9314349		✘	✘	✘
rs10787010		✘	✘	✘
rs10787011	✘		✘	✘
rs11191692	✘	✘	✘	✘
rs11601726				✘
rs12209631	✘	✘	✘	✘
rs16924159	✘		✘	✘
计数	14	16	17	22

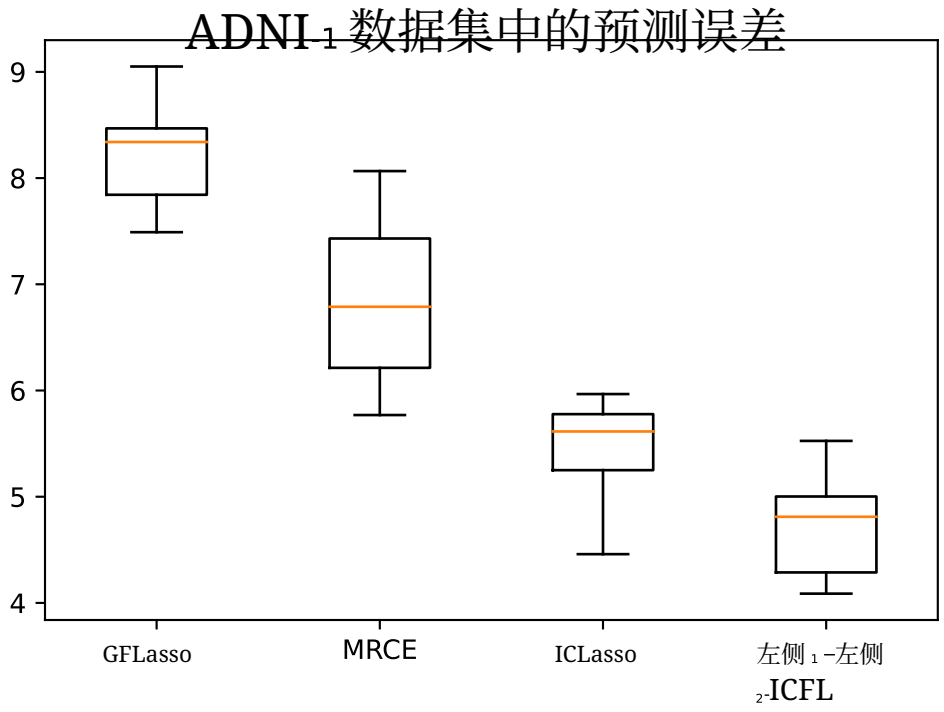
6. 结论。本文提出了一种基于 l_1与 l_2 范数之差的图形套索，为 B的稀疏性引入了一种新的惩罚项。基于两个重要假设，我们联合估计回归系数矩阵 B 和输出结构Θ。与 ICLasso 类似，该优化问题可利用现有算法求解。通过合成数据集，我们证明$l_{1\text{-}2}$-GLasso 在恢复 eQTL 关联和基因网络结构方面优于其他模型。真实数据集上的结果也显示了$l_{1\text{-}2}$-GLasso 的优越性，并验证了模型的假设。此外，我们还可以对 B 和 Θ 采用其他惩罚范数，例如$l_{\frac{1}{2}}$ 惩罚，它与 l_0惩罚更为接近。未来的工作将致力于提高求解算法的效率，并分解大规模问题。该模型的应用领域也可以扩展到金融数据。

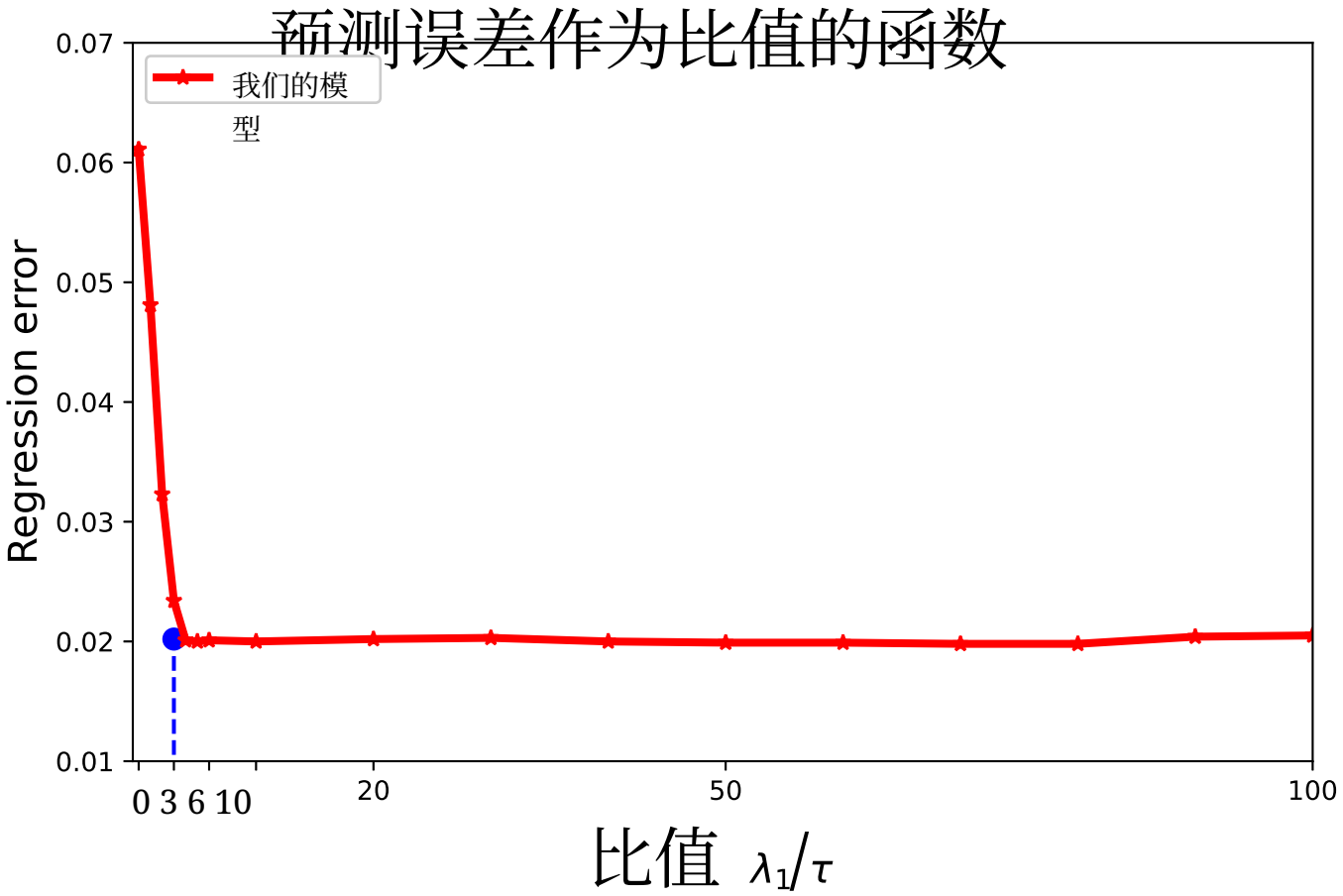
数据可用性。本文分析的数据可在 R 包“bgsmtr”中获取。在 R 控制台中运行：1.data(bgsmtr-example-data) 2.str(bgsmtr-example-data) 3.单核苷酸多态性 <- t(bgsmtr-example-data-SNP-data) 4.脑测量 <- t(bgsmtr-example-data-Brain-Measures)

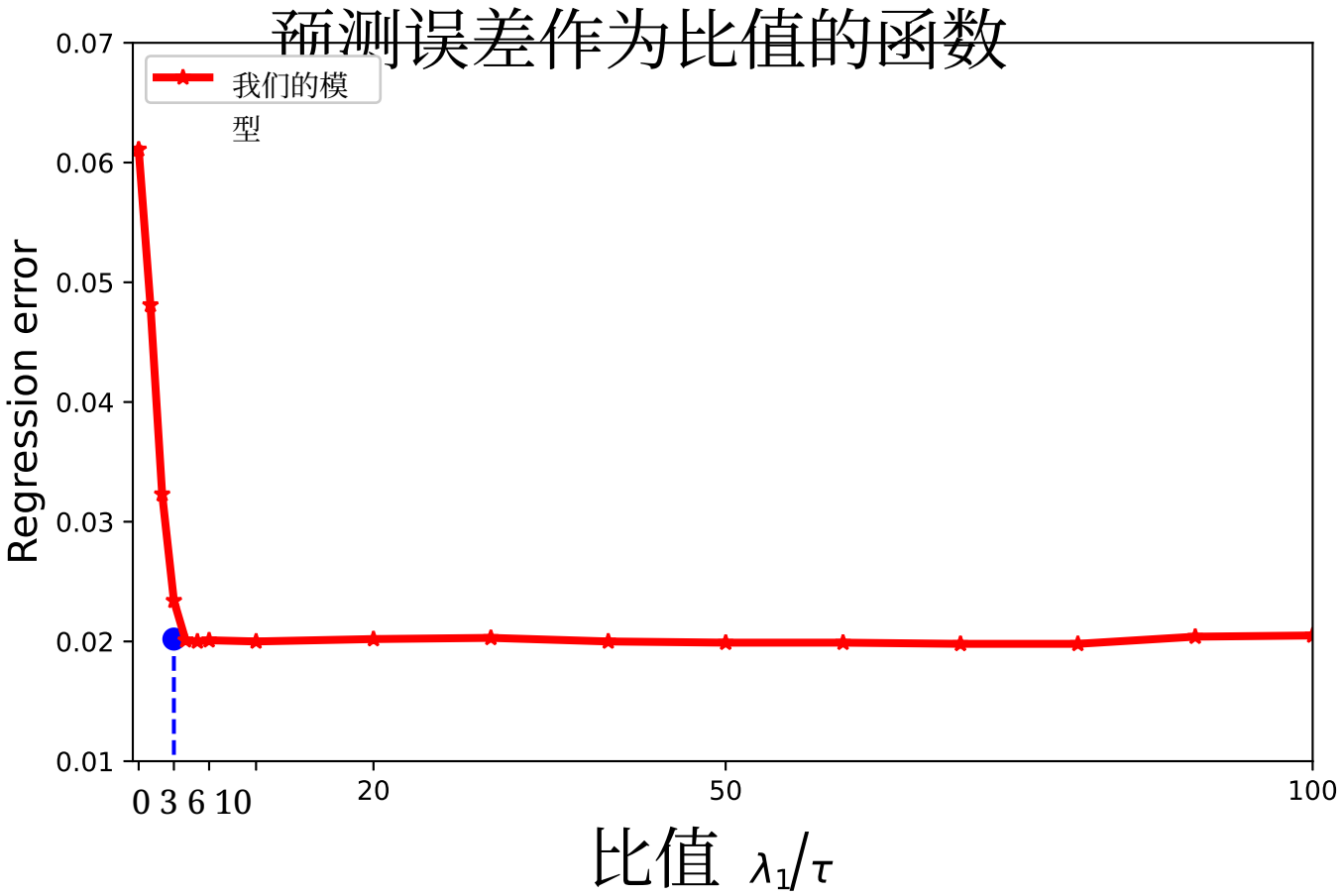
参考文献

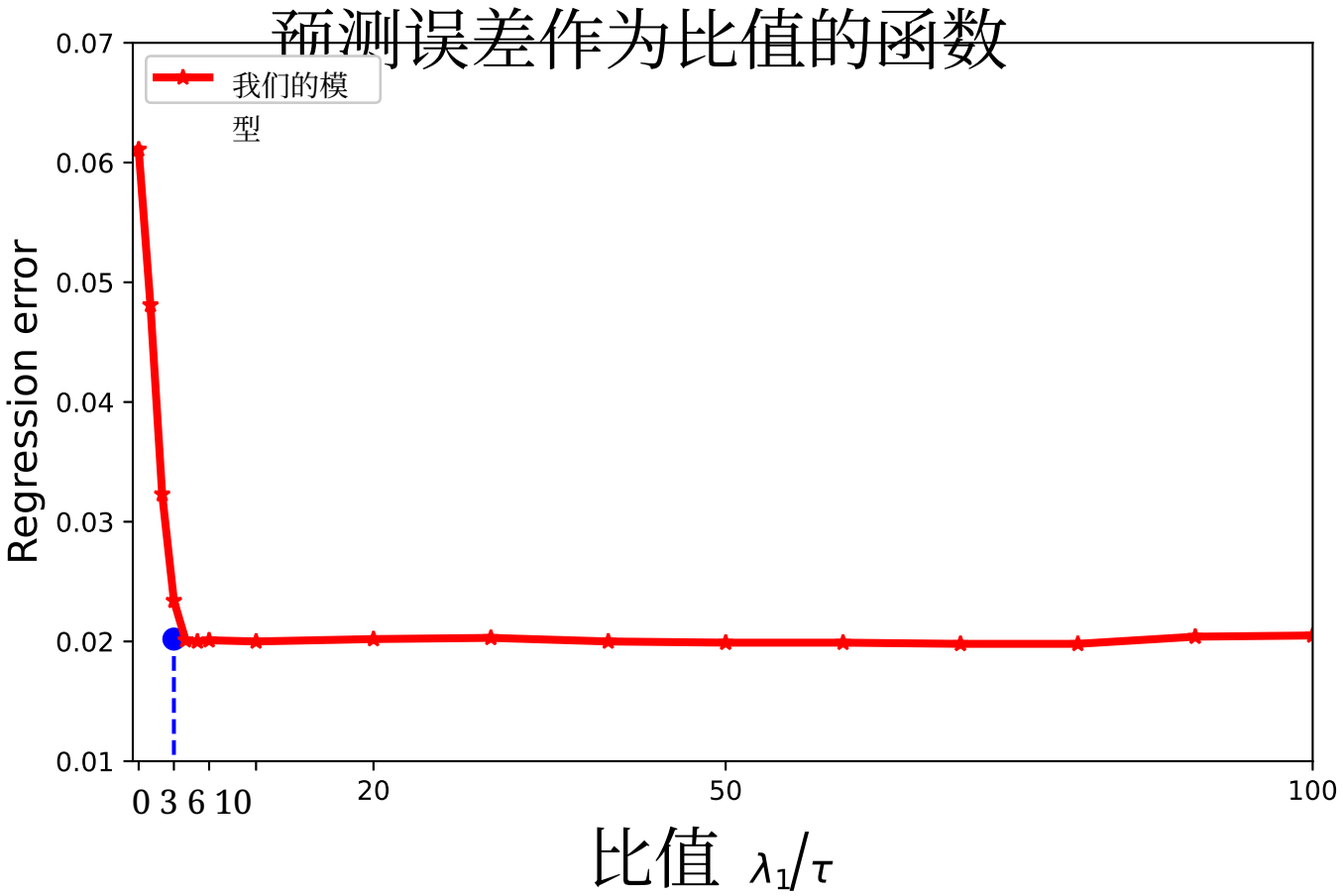
BERTSIMAS, D. 和 VAN PARYS, B.（2020年）。稀疏高维回归：精确、可扩展的算法与相变。《统计年刊》48 300–323。CHUN, H.、CHEN, M.、LI, B. 和 ZHAO, H.（2013年）。基于多源基因组数据的联合条件高斯图模型。《遗传学前沿》4 294。ESSER, E.、LOU, Y. 和 XIN, J.（2013年）。一种求解非负最小二乘问题中结构化稀疏解的方法及其应用。《SIAM 成像科学杂志》6 2010–2046。FRIEDMAN, J.、HASTIE, T. 和 TIBSHIRANI, R.（2008年）。利用图形套索进行稀疏逆协方差估计。《生物统计学》9 432–441。

GARDNER, T. S. 和 FAITH, J. J.（2005年）。转录调控网络的逆向工程。《生命物理评论》2 65–88。GREENLAW, K.、SZEFER, E.、GRAHAM, J.、LESPERANCE, M.、NATHOO, F. S. 和 INITIATIVE, A. D. N.（2017年）。用于影像遗传学的贝叶斯组稀疏多任务回归模型。《生物信息学》33 2513–2522。HOYER, P. O.（2004年）。带稀疏性约束的非负矩阵分解。《机器学习研究杂志》5。KIM, S. 和 XING, E. P.（2009年）。定量性状网络中相关基因组关联的统计估计。《公共科学图书馆·遗传学》5 e1000587。KIM, S. 和 XING, E. P.（2012年）。用于结构化稀疏多响应回归的树引导组套索及其在表达数量性状位点定位中的应用。《应用统计年刊》6 1095–1117。LI, B.、CHUN, H. 和 ZHAO, H.（2012年）。条件图模型的稀疏估计及其在基因网络中的应用。《美国统计学会会刊》107 152–167。LI, H. 和 GUI, J.（2006年）。用于稀疏高斯浓度图的梯度导向正则化及其在遗传网络推断中的应用。《生物统计学》7 302–317。LIU, C.、CHEN, Q.、ZHOU, B. 和 LI, H.（2016年）。基于 L_1-和 L_2-范数联合正则化的稀疏信号重构方案。《工程数学问题》2016。MARBACH, D.、PRILL, R. J.、SCHAFFTER, T.、MATTIUSSI, C.、FLOREANO, D. 和 STOLOVITZKY, G.（2010年）。揭示基因网络推断方法的优势与不足。《美国国家科学院院刊》107 6286–6291。MARCHETTI-BOWICK, M.、YU, Y.、WU, W. 和 XING, E. P.（2019年）。用于联合估计 eQTL 关联与基因网络结构的惩罚回归模型。《应用统计年刊》13 248–270。MEINSHAUSEN, N. 和 BÜHLMANN, P.（2006年）。基于套索回归的高维图与变量选择。《统计年刊》34 1436–1462。MICHAELSON, J. J.、ALBERTS, R.、SCHUGHART, K. 和 BEYER, A.（2010年）。基于数据评估表达数量性状位点定位方法。《BMC基因组学》11 1–16。PENG, J.、WANG, P.、ZHOU, N. 和 ZHU, J.（2009年）。利用联合稀疏回归模型进行偏相关估计。《美国统计学会会刊》104 735–746。ROCKMAN, M. V. 和 KRUGLYAK, L.（2006 7 862 872）。全局基因表达的遗传学。《自然综述：遗传学》–。ROTHMAN, A. J.、LEVINA, E. 和 ZHU, J.（2010年）。带协方差估计的稀疏多元回归。《计算与图形统计学杂志》19 947–962。SCHÄFER, J. 和 STRIMMER, K.（2005年）。用于大规模协方差矩阵估计的收缩方法及其对功能基因组学的启示。《遗传学与分子生物学中的统计应用》4。SEGAL, E.、FRIEDMAN, N.、KAMINSKI, N.、REGEV, A. 和 KOLLER, D.（2005年）。从特征到模型：利用微阵列理解癌症。《自然遗传学》37 S38–S45。TIBSHIRANI, R.（1996年）。通过套索回归实现回归收缩与变量选择。《英国皇家统计学会杂志：B辑（统计方法学）》58 267–288。VINGA, S.（2021年）。用于分析高维组学数据的结构化稀疏正则化。《生物信息学简报》22 77–87。WANG, H.、NIE, F.、HUANG, H.、KIM, S.、NHO, K.、RISACHER, S. L.、SAYKIN, A. J.、SHEN, L. 和 INITIATIVE, A. D. N.（2012年）。通过组稀疏多任务回归和特征选择识别数量性状位点：ADNI队列的影像遗传学研究。《生物信息学》28 229–237。WANG, L.、ZHOU, H.、WANG, Y.、YU, B.、ZHANG, Y.、LIU, W. 和 CHEN, Y.（2019年）。基于 L_{1-2} 最小化的三参数叠前地震反演。《地球物理学》84 R753–R766。YIN, P.、ESSER, E. 和 XIN, J.（2014年）。l_1 范数与 l_2 范数的比值和差值，以及基于相干字典的稀疏表示。《信息与系统通信》14 87–109。YIN, J. 和 LI, H.（2011年）。用于遗传基因组学数据分析的稀疏条件高斯图模型。《应用统计年刊》5 2630。YIN, P.、LOU, Y.、HE, Q. 和 XIN, J.（2015年）。压缩感知中的 l_{1-2} 最小化。《SIAM科学计算杂志》37 A536–A563。YU, Y.-L.（2013年）。利用近端平均实现更好的近似与更快的算法。《神经信息处理系统进展》26。YUAN, M. 和 LIN, Y.（2006年）。含分组变量的回归模型选择与估计。《英国皇家统计学会杂志：B辑（统计方法学）》68 49–67。


ADNI1数据集中的预测误差
9
8
7
6
5
4
GFLasso
MRCE
ICLasso
左侧1-左侧2-ICFL


预测误差作为比值的函数
我们的模型
Regression error
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0 3 6 10
20
50
100
比值 λ_1/τ


预测误差作为比值的函数
我们的模型
Regression error
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0 3 6 10
20
50
100
比值 λ_1/τ


预测误差作为比值的函数
我们的模型
Regression error
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0 3 6 10
20
50
100
比值 λ_1/τ